I.7. Statistical Quality Control (SQC)

All efforts aimed at making optimal use of observations in an atmospheric analysis will fail if the wrong data are used. Inclusion of a few corrupted data, or exclusion of a few accurate observations that happen to disagree with a model forecast, can make the difference between a good and a bad forecast. GEOS Terra uses many different types of observations, originating from a various instruments and obtained by multiple communication channels. Some—but not all—of the data have been subject to quality control checks at some point during data preprocessing. Once observations are ingested into the DAS in a uniform format, they can be compared with each other as well as with the global *a priori* estimate of the atmospheric state produced by the DAS.

The on-line Statistical Quality Control (SQC) system attempts to identify observations that are likely to be contaminated by gross errors. The algorithms involve statistical tests of the actual data against assumptions about their expected errors and about GCM forecast errors. Essentially, a local statistical analysis is performed for each outlier observation, i.e., for each observation that differs significantly from the short-term forecast produced by the GCM. If this analysis indicates that the observation is inconsistent with surrounding data, then that observation is marked for rejection.

The SQC encompasses a *background check*, a *buddy check*, a *wind check*, and a *profile check*, each of which is described below. All checks are formulated in terms of the observed-minus-forecast residuals (O-F) rather than the observations themselves. All checks potentially modify the quality control marks associated with the observations, but leave all other data attributes unchanged. The background check and buddy check involve the forecast and observation error variances for the quantities being tested, which are prescribed in the global analysis system.

I.7.1. Statistical aspects

The SQC algorithms operate on the vector of observed-minus-forecast residuals v defined by

$$v = w^o - f(\mathcal{I}w^f), \tag{7}$$

where w^o is the vector of observations, w^f is the forecast vector, f is the observation operator, and $\mathcal{I}$ is the linear operator which interpolates state variables from model grid points to observation locations. The observation operator maps model variables to observables. For remotely sensed radiances, for example, the function f represents a radiative transfer model. It is simply the identity for conventional, *in situ* observations of model variables.

The SQC attempts to identify corrupt data based on statistical expectations. This requires

knowledge of the covariance S of the observed-minus-forecast residuals, defined by

$$S_{ij} = \langle v_i v_j \rangle, \tag{8}$$

with i, j indicating location. In general these covariances are poorly known, but a rough estimate is available from the global analysis system. It follows from (7) that

$$S \approx F\mathcal{I}P^f\mathcal{I}^T F^T + R, \tag{9}$$

where F is the linearized observation operator

$$F = \left.\frac{\partial f}{\partial w}\right|_{w=w^f}, \tag{10}$$

and P^f, R are the covariances of forecast and observation errors, respectively. Equation (9) would be exact if forecast and observation errors were entirely independent (they are not, since both types of errors depend on the true state) and if all observation operators were linear.

Specification of reasonably accurate error covariances is crucial to the quality of a statistical analysis. We therefore assume that the right-hand side of (9), as prescribed by the global analysis system, provides some useful information about the residual error covariances. Accordingly, prescribed error statistics are used to define tolerances for the background check, whose main purpose is to mark outlier observations for subsequent reexamination in the buddy check. However, since actual errors depend on many unknown model defects and other intangibles, covariance specifications in operational data assimilation systems cannot be relied upon to accurately describe error characteristics in all situations at all times. In particular, during extreme events—when quality control decisions become especially important—the covariances as prescribed by the global analysis system are almost certainly inadequate. Thus, a key aspect of the SQC is the attempt to adjust the prescribed error statistics based on actual data. This adjustment takes place during the buddy check, before a final accept/reject decision is reached for an outlier observation.

I.7.2. The background check

The background check tests each single observation against a background estimate, which is simply the 6-hour model forecast interpolated to the time and location of the observation. If the discrepancy is extremely large then the observation is rejected outright. If the discrepancy is large, but not extremely large, then the observation is marked as an outlier, to be reexamined in the buddy check. The tolerances for the background check are defined in terms of standard deviations obtained from the error statistics as prescribed by the global analysis system.

The algorithm is as follows:

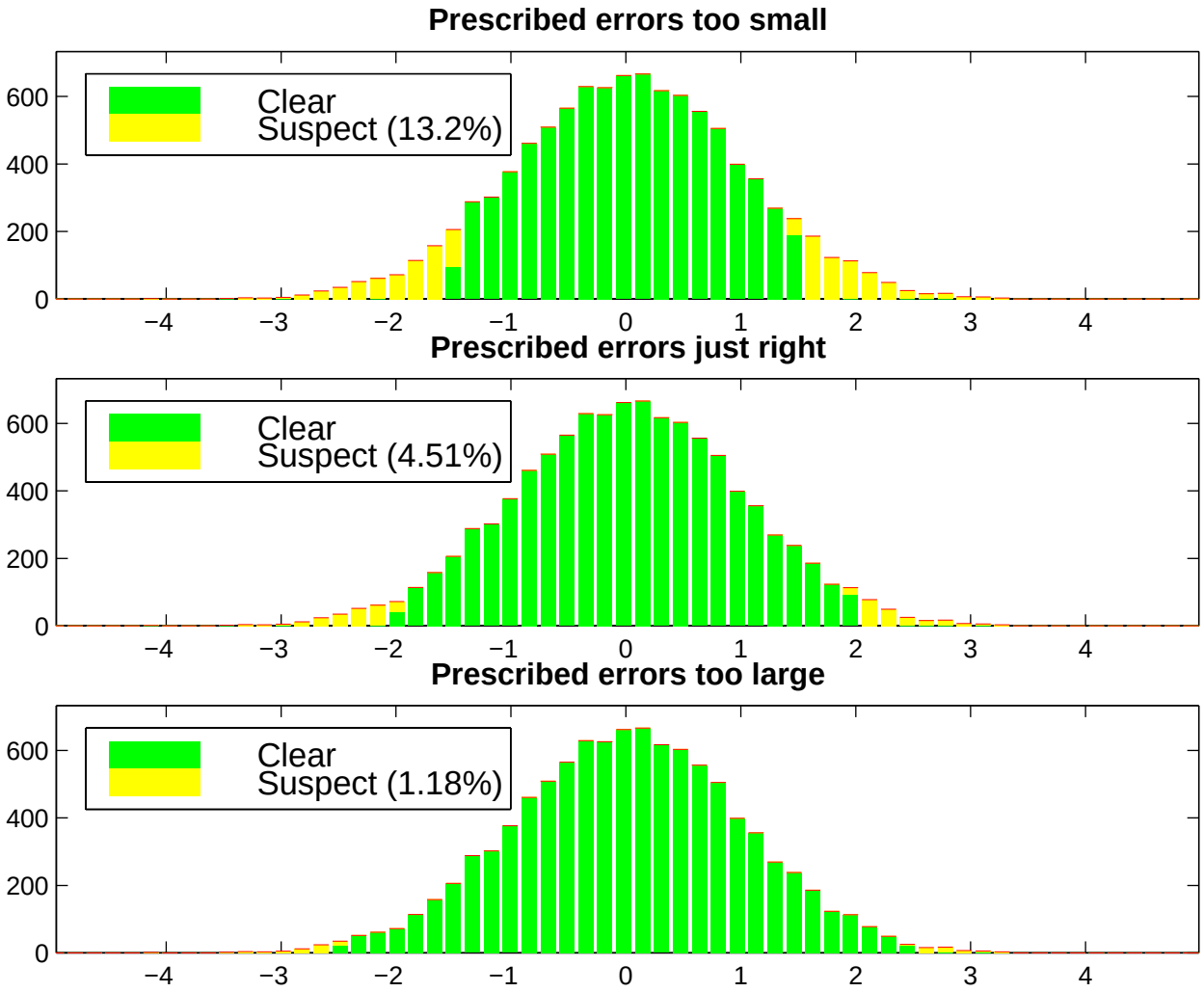

Figure 9: Illustration of the relationship between the rate at which the background check marks observations as outliers and the prescribed error statistics, for normally distributed errors. The yellow tails of the histograms correspond to observations marked as outliers.

For each observation w_i^o:

$$\begin{array}{ll} \text{mark } w_i^o \text{ as an } \textbf{outlier} \text{ if} & |v_i| > \tau_o \sigma_i \\ \text{mark } w_i^o \text{ as } \textbf{excluded} \text{ if} & |v_i| > \tau_x \sigma_i \end{array}$$

Here $\sigma_i = \sqrt{S_{ii}}$, and τ_s, τ_x are prescribed non-dimensional tolerance parameters. Typically we take $\tau_o = 2, \tau_x = 10$. The rate at which the background check produces suspect marks presents a useful check on the accuracy of the prescribed error statistics. If the forecast and observation error variances are correctly tuned, and if the errors are roughly normally

distributed, then the suspect rate can be predicted. For example, when $\tau_o = 2$, the rate should be about 4.5%. If the actual suspect rate is larger (smaller), then the prescribed error variance is too small (large). This is illustrated in Fig. 9. Monitoring the background check failure rates for specific instruments has, in a number of cases, led to adjustments of observation error statistics in GEOS Terra.

I.7.3. The buddy check

The buddy check is applied to a subset of observations which are considered suspect, either because they were identified as an outlier by the background check, or because they were marked as suspect during the preprocessing stage. The buddy check attempts to predict the value of a suspect observation from nearby non-suspect observations (the *buddies*.) If the predicted value is in reasonable agreement with the observation, then the observation is no longer considered suspect. If a sufficient number of buddies is available, then the tolerance for the buddy check is adjusted based on a local estimate of O-F standard deviations. Once all suspect observations have been tested, the entire process is repeated for all observations that are still considered suspect. The process stops when the set of suspects no longer changes: all remaining suspects are then rejected.

The buddy check initially labels observations as suspect based on their quality control history. A single iteration of the algorithm is as follows:

For each suspect observation w_j^o:

1. Define the set of buddies:

   Nearby non-suspect observations of the same data type as w_j^o are ranked according to the scalar weight that each would receive in an optimal univariate statistical analysis at the location of w_j^o. The buddies are simply the n highest ranking of these, where n is a configuration parameter. Typically we take $n = 50$.

2. Predict the value of the suspect observation based on its buddies:

   Using the weights determined in the previous step, the weighted average $v_j^\star$ of the v_i associated with the buddies provides the optimal univariate analysis of the buddies at the location of w_j^o.

3. Adjust the prescribed estimate of the local O-F standard deviation:

   If $\widehat{\sigma}_j^2$ is the sample variance of the v_i associated with the buddies, the prescribed variance σ_j^2 is adjusted according to

   $$(\sigma_j^\star)^2 = (n^\star \sigma_j^2 + n\widehat{\sigma}_j^2)/(n^\star + n) \tag{11}$$

   where $n^\star$ is a configuration parameter. Typically we take $n^\star = 25$.

4. Reevaluate the status of w_j^o:

   Change the status of w_j^o to non-suspect if

$$|v_j - v_j^\star| < \tau_b \sigma_j^\star \tag{12}$$

   where τ_b is a prescribed non-dimensional tolerance parameter. Typically we take $\tau_b = 3$.

These steps are repeated until no further observations change status. At that point, any remaining suspect observations are marked for rejection.

The adaptive nature of the buddy check has two important consequences. First, the final quality control decisions are not very sensitive to the prescribed error statistics in the global analysis system. We have verified this experimentally by varying the tolerance parameter τ_o of the background check. It was found that the final accept/reject status of observations is not very sensitive to the background check failure rate, as long as this rate is roughly between 1% and 10%. This insensitivity to the prescribed statistics is a major practical advantage, since (1) these statistics are not very reliable and (2) the SQC algorithms do not require retuning each time the prescribed statistics in the global analysis change.

The second consequence of adjusting rejection limits on the fly based on the local variability of surrounding data is that the buddy check becomes increasingly tolerant in synoptically active situations (and, conversely, more stringent when the flow is smooth). This is best illustrated by an example, in which we contrast the results of a nonadaptive buddy check against those of the adaptive buddy check. Figure 10 shows two maps with quality control marks for zonal wind observations (obtained from aircraft and rawinsonde reports) over North America at or near 200hPa, on January 14 1998. The top panel shows rejections (indicated by red marks) by a non-adaptive buddy check, based on tolerances derived from prescribed statistics. Yellow marks indicate data that were marked as outliers by the background check, but which passed the buddy check. The lower panel shows rejections by the adaptive buddy check. Tolerances are increased due to greater variability than implied by the prescribed statistics, resulting in the acceptance of several additional outlier observations. The effect on the wind analysis (not shown) is to increase wind speeds by about $5m/s$ in some places.

I.7.4. The wind check

This check is applied to all u-wind and v-wind data to make sure that wind components pass the quality control in pairs. The algorithm determines whether two wind components are paired (i.e., whether they originate from the same report) by matching their location attributes, instrument type, and sounding index.

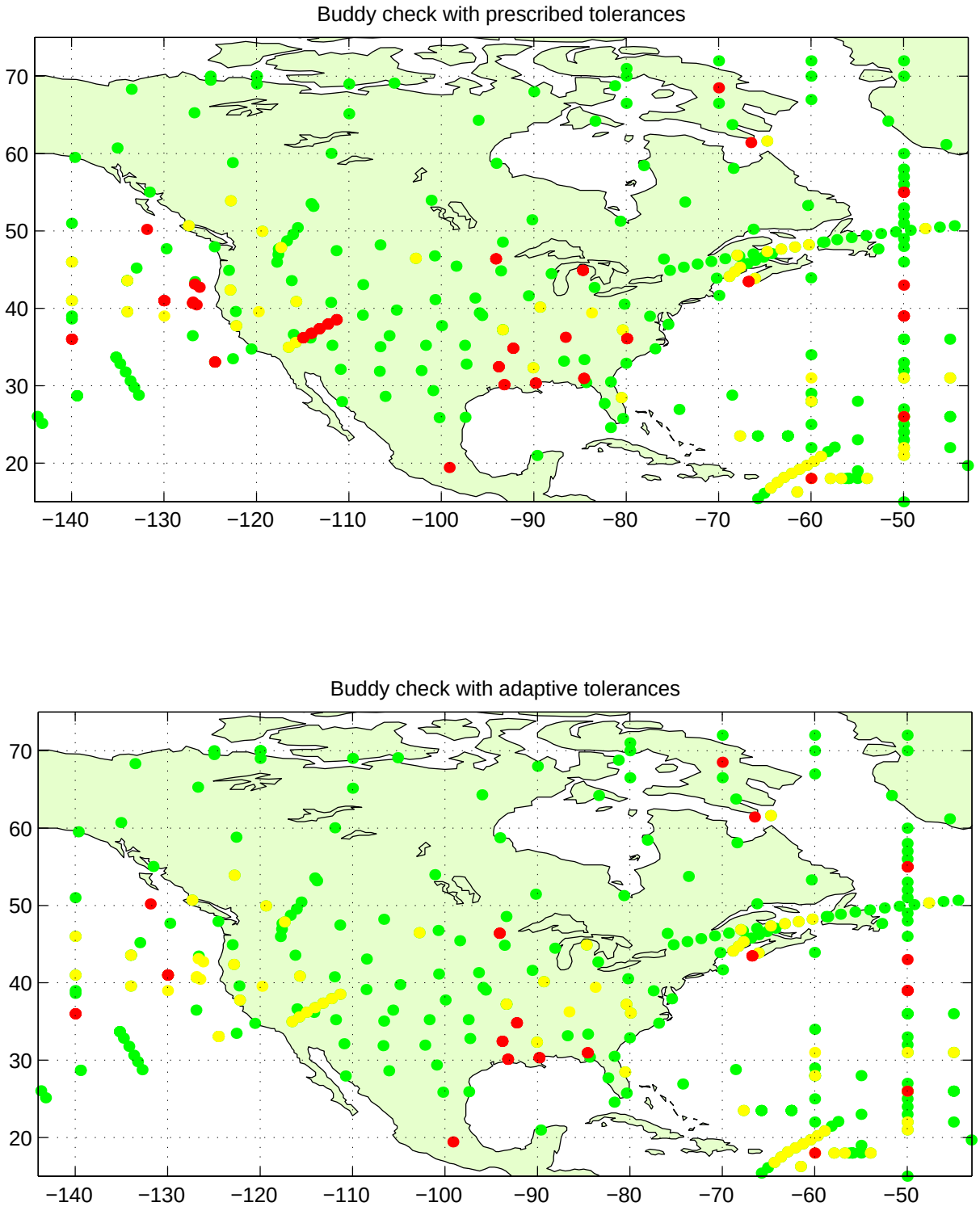

Figure 10: Quality control decisions for zonal wind observations at 200hPa on January 14 1998, using a non-adaptive buddy check (top) and adaptive buddy check (bottom).

I.7.5. The profile check

This check eliminates an entire vertical sounding in case any of the data from that sounding are marked for exclusion. It is applied to selected data types only. Currently the profile check is used for TOVS height retrievals only. For example, if the buddy check rejects a TOVS height observation at 10hPa, then the entire sounding is marked for rejection.

I.7.6. Special treatment of moisture observations

The analysed moisture field in GEOS Terra is water vapor mixing ratio, which is highly variable in space and time. This causes difficulties for the buddy check, which presumes that the field is spatially coherent on the scales resolved by the observing network. Experience has shown that a buddy check applied to water vapor mixing ratio observations (or, equivalently, specific humidity) tends to reject too many of them, unless the tolerances are relaxed to a point where the quality control becomes almost completely inactive. This is obviously not acceptable, unless preprocessing quality control is completely reliable.

To remedy this situation, the statistical tests (background check and buddy check) in the SQC are applied to relative humidity residuals. These residuals are computed in two ways: first, using observed mixing ratios and observed temperatures, and second, using observed mixing ratios and model-predicted temperatures. This prevent the situation in which a relative humidity looks good even though both mixing ratio and temperature are corrupt. The tests are applied in sequence to both types of residuals, and an observation passes QC only if none of the tests fail.